The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^{2}HCAI)

Towards a Deeper Understanding of Concept Bottleneck Models Through End-to-End Explanation

Jack Furby[1] **Daniel Cunnington**[2] **Dave Braines**[2] **Alun Preece**[1]

[1] Cardiff University, UK [2] IBM Research Europe
furbyjl@cardiff.ac.uk

Abstract

Concept Bottleneck Models (CBMs) first map raw input(s) to a vector of human-defined concepts, before using this vector to predict a final classification. We might therefore expect CBMs capable of predicting concepts based on distinct regions of an input. In doing so, this would support human interpretation when generating explanations of the model's outputs to visualise input features corresponding to concepts. The contribution of this paper is threefold: Firstly, we expand on existing literature by looking at relevance both from the input to the concept vector, confirming that relevance is distributed among the input features, and from the concept vector to the final classification where, for the most part, the final classification is made using concepts predicted as present. Secondly, we report a quantitative evaluation to measure the distance between the maximum input feature relevance and the ground truth location; we perform this with the techniques, Layer-wise Relevance Propagation (LRP), Integrated Gradients (IG) and a baseline gradient approach, finding LRP has a lower average distance than IG. Thirdly, we propose using the proportion of relevance as a measurement for explaining concept importance.

Introduction

Humans build mental models which are internal representations of an object's internal mechanics. These are used to predict an object's future states and aid in interactions (Johnson-Laird 1986; Craik 1943; Halasz and Moran 1983; Norman 1983). If a human is unable to build an accurate representation of a Deep Neural Network (DNN) decision boundaries, the human could be misled to either accept misclassifications or disregard its output entirely. Explainable artificial intelligence (XAI) techniques aim to aid humans in building mental models of DNNs. One XAI technique is to use *saliency maps*, a tool which highlights features of an input which were relevant to a prediction.

Concept Bottleneck Models (CBMs) (Koh et al. 2020) seek to enable richer human-machine interaction by training a DNN to predict a vector of human-defined concepts before using this vector to predict a final classification. CBMs have the potential of performing a task in a similar way to humans, whilst also enabling interpretability of their learned representations. For instance, to identify a bird, a human may first recognise parts of the bird such as the colour and size, before using this information for bird identification. As the final classification uses the predicted concept vector, a human user will be able to modify the concept vector, referred to as *intervening*, and add or remove concepts to inspect changes to the final classification.

Despite the concept vector output, CBMs are unable to explain what input features lead to concept predictions, or which concepts contribute to the final classification. An XAI study for CBMs (Margeloiu et al. 2021) used saliency maps and suggests that CBMs do not learn concepts as humans would expect, but instead attribute relevance to the entire input, and not to distinct regions. The authors however only looked at saliency maps for concepts and not the final classification and did not indicate what the models may be learning for concept predictions. In this paper we position XAI, and in particular saliency maps, as a technique to present the relevancy behind a CBM prediction, both concept and final classification and thus make the model reasoning accessible to a human.

Our research focuses on producing explanations for CBMs targeted for use by domain experts. For this reason, we use CBMs with Layer-wise Relevance Propagation (LRP) (Bach et al. 2015) to explore relevancy attributed, from the final classification to the concept vector and from the concept vector to the input as relevance can be attributed to groups of input features instead of on a pixel-by-pixel basis (Samek et al. 2021). Despite the primary use of LRP in this paper, the techniques described apply to other gradient-based attribution methods such as Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017) which we also used to compare concept vector to input saliency maps. Our code and additional results are publicly available[1].

Background

Concept Bottleneck Models

CBMs can be given in the form $f(g(x))$ where the function g refers to the prediction of concepts $\hat{c}$ using the input x and the function f is the prediction of the final classification y with the input $\hat{c}$.

Given the training set $\{x^{(i)}, y^{(i)}, c^{(i)}\}_{i=1}^{n}$ where we are provided with a set of inputs $x \in \mathbb{R}^d$, corresponding targets

[1] Code and results: https://github.com/JackFurby/explainable-concept-bottleneck-models

$y \in \mathcal{Y}$ and vectors of k concepts $c \in \mathbb{R}^k$. A CBM maps the input space to the concept space $g : \mathbb{R}^d \to \mathbb{R}^k$ and maps concepts to final targets $f : \mathbb{R}^k \to \mathcal{Y}$. This is such that the final classification is made using only concept predictions.

CBMs can be trained in three ways: *independent*, *sequential* and *joint*. With independent training, each model part is trained separately, whereas the sequential method trains the model parts one after another and joint trains them together in an end-to-end fashion.

In this paper, we refer to the model part predicting c as $x \to c$ and the model part predicting y as $c \to y$.

Layer-wise Relevance Propagation

LRP (Bach et al. 2015) is an explanation technique that propagates a prediction backwards through a network until a defined layer or input is reached, following a set of rules. Primarily, the network output is conserved and is only redistributed to the neurons of the previous layer. The total value propagated does not change. The use of alternative rules adds flexibility for LRP implementation (Montavon et al. 2019). These include the basic rule (LRP-0), Epsilon rule (LRP-ϵ) and Alpha Beta Rule (LRP-$\alpha\beta$) (Bach et al. 2015). Rules can be applied in a composite manner to overcome the shortcomings of any single rule.

LRP is considered to have an advantage over the related IG method (Sundararajan, Taly, and Yan 2017) in that IG tends to produce very fine-grained pixel-wise mappings whereas LRP tends to group relevance to features from the input (Samek et al. 2021). As we are interested in mapping relevance attributed to concepts, and concepts occupy distinct regions in input images, this makes LRP an appealing choice.

Setup

Models and Dataset

Our models were trained with the same modifications of the CUB-200 2011 (CUB) dataset (Wah et al. 2011) as (Koh et al. 2020) which used class-level concepts with 11,788 bird images covering 200 classifications and 112 concepts.

We trained four models, one for each of the independent and sequential methods and two for the joint method (with and without a sigmoid activation between the two model parts). For the $x \to c$ model part, we used a VGG-16 architecture with batch normalisation (Simonyan and Zisserman 2015), pre-trained on ImageNet. The $c \to y$ model part is a single fully connected layer. Model performance on the test dataset is shown in Table 1.

LRP Configuration

For the $x \to c$ model we use similar LRP rules to (Montavon et al. 2019). These rules are: LRP-$\alpha\beta$, where $\alpha = 1$ and $\beta = 0$, for the first seven convolutional layers of the model from the input, LRP-ϵ for the next six convolutional layers and LRP-0 for the top three linear layers. LRP-0 is used for the $c \to y$ model.

We used a method detailed by (Taylor et al. 2020), changing modality for concepts, to calculate the proportion of relevance for each concept, and thus the percentage each concept contributed towards the final classification. This is possible because LRP conserves relevancy. As each concept is a single value we do not need to account for imbalance in concept proportions.

By calculating the contribution of concepts for the final classification, a human may be able to focus on the most influential concepts to a final classification and, if intervention is required, which concepts they may wish to intervene on.

Training method	Classification accuracy	Concept accuracy
Independent	77.51%	96.85%
Sequential	75.35%	96.85%
Joint-without-sigmoid	78.75%	96.12%
Joint-with-sigmoid	75.35%	94.87%

Table 1: Models final classification top-1 accuracy and concept binary accuracy

Results

Figure 1 shows the relevance from the concept layer back to the input for a range of concepts which a human would expect to map to distinct regions of the input. Regardless of the training method used, the saliency maps indicate that the models have not learned a mapping of distinct regions in the input to concepts. Relevancy is generally distributed over the entire bird although an observation with our models is the eyes of the bird appears to be the most common feature to be highly positive or negatively relevant.

Concepts with similar predictions also appear to share similar saliency maps. This is evident in Figure 1 with the *independent and sequential* models and concepts *has_crown_color::brown* and *has_wing_shape::pointed-wings* which have a predicted concept value of 0.9973 and 0.9980 respectively to four decimal places. For the *joint-without-sigmoid model*, *has_back_color::brown* has a predicted concept value of 0.9918 and *has_breast_pattern::solid* has a predicted concept value of 0.9975. The similarity between saliency maps likely means that each model has learned the same input features, can accurately predict different concepts.

Our results confirm CBMs trained on the CUB dataset do not learn distinct regions from the input to concepts, as (Margeloiu et al. 2021) showed. This is likely due to the training data or training methods not constraining the model to do so. Like regular bottleneck models (Grezl et al. 2007), CBMs will typically only keep the most important input features, in this case, to fit the concept vector, but leave the CBM to select which input features to use. In addition, by using class-level concepts the model learns the concept vector but not if a concept is present and visible in a given sample. Koh et al. (2020) version of CUB also has incorrect concepts. For example, class *Mallard* has the same concept vector for males and females despite the visual differences between them. If the dataset instead had instance-level con-

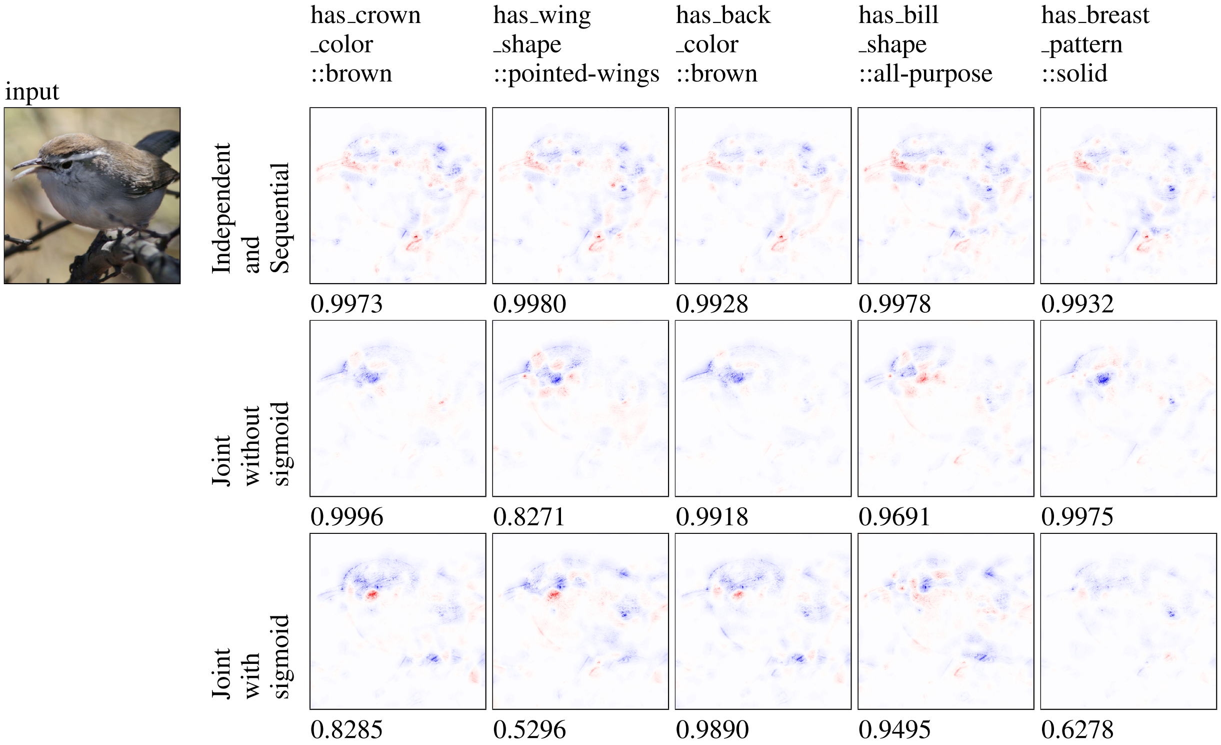


Figure 1: Concept saliency maps for the input of a Bewick Wren image and correctly predicted concepts. Positive relevance is shown in red, negative relevance is shown in blue and the predicted concept value to four decimal places, and sigmoid applied, below each saliency map. In general, relevance does not map to input features that a human would associate each concept with.

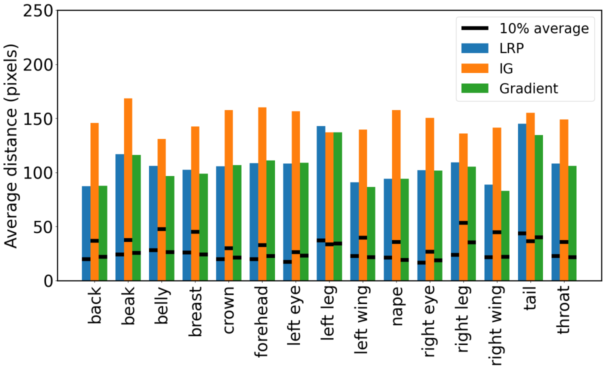
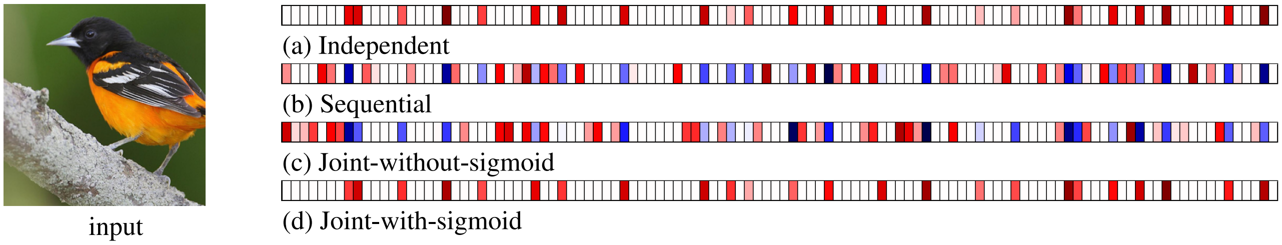


Figure 2: Distance pointing game results comparing LRP, IG and a baseline gradient method. LRP and gradient has a shorter average distance for most bird parts compared to IG. This remains the same for when averaging the shortest 10% of distances.

cepts, where each sample has its own concept vector only showing concepts present, we may see concept predictions that are closer to how humans would perceive them.

We also evaluated IG with a SmoothGrad noise tunnel (Smilkov et al. 2017) using a batch size of 25 and a standard deviation of 0.2, similar to (Margeloiu et al. 2021), using a quantitative evaluation method in comparison with LRP and a baseline gradient method (Simonyan, Vedaldi, and Zisserman 2014). For our evaluation we modified the pointing game (Zhang et al. 2018) which counts hits and misses whether the most salient point of a given saliency map was within a defined region, the ground truth, resulting in an accuracy measurement. Our version, called distance pointing game, averages the distance between the most salient point of a saliency map and the ground truth point. (This was necessary because CUB does not provide bird part bounding boxes.) Our technique does not replace the pointing game, but instead, it satisfies a different situation; when you have ground truth points. By using our evaluation technique, we can quantify whether a saliency technique for a given model's output is primarily focusing on a ground truth point. We can also rank saliency techniques or models, which enables us to analyse further.

We measured the average distance using our independent model, due to that model having the highest concept accuracy, and the validation dataset split. Results are shown in Figure 2. IG has around a 3rd higher average distance compared to both LRP and the baseline gradient for most bird parts while LRP and the baseline have similar average distances. To remove noisy saliency maps we also show the average of the shortest 10% of distances which follows the same story as the overall distance averages As LRP with our


input

(a) Independent

(b) Sequential

(c) Joint-without-sigmoid

(d) Joint-with-sigmoid

Figure 3: Final classification saliency maps for a correctly predicted Baltimore Oriole input. Each vector has 112 segments, one for each concept input. Positive relevance is shown in red and negative relevance is shown in blue. The independent and joint-with-sigmoid models only apply positive relevance to concept predicted as present. The joint-without-sigmoid and sequential models apply positive relevance to concepts predicted as not present and negative relevance to concepts predicted as present.

rule setup groups relevance to input regions, while IG applies relevance on a pixel-by-pixel basis (Samek et al. 2021), LRP saliency maps are filtering out noisy relevance from the input image. However, the average distance hovers around 100 pixels away from the ground truth point with LRP and, considering the input images are 299 by 299 pixels in size, this could still fall outside of the concept in the input image, adding to what we observed in Figure 1 with relevance generally covering the entire bird.

Applying LRP to the $c \rightarrow y$ model, unlike the $x \rightarrow c$ model, we can produce saliency maps with closer alignment to human decision making. Figure 3 presents LRP saliency maps for the $c \rightarrow y$ models where they show the independent and joint-with-sigmoid training methods have learned a mapping from predicted concepts to classification that exclusively uses concepts predicted as present for the final classification. Both the sequential model and joint-without-sigmoid model applies relevance to concepts predicted as present and not present but with concept predicted as not present having positive relevance, while concepts predicted as present have negative relevance. These models appear to have learned a mapping from the concept vector prioritising the absence of concepts rather than the presence of them. Relevance is not flipped for all samples in the test dataset for these two models although it occurs often enough for it to be noteworthy. Saying for certain why the model appears to apply positive relevance to concepts that are not present and, if this is the general case of CBMs, or just models using the CUB dataset, remains an open question.

As discussed earlier, LRP enables us to calculate the contribution of each predicted concept to the final classification. For the same input as used in Figure 3, the top three concepts contributing to the final class predicted with the independent model are as follows: *has_upperparts_color::white* at 6.04%, *has_primary_color::yellow* at 5.83%, and *has_tail_pattern::multi-colored* at 5.39% with a total of 38 concepts contributing to the final classification. By calculating the concept contributions we are revealing the $c \rightarrow y$ model part reasoning towards the final classification such that a human can take this into their own decision-making when interacting with a CBM.

Conclusion and Future Work

This paper evaluates CBMs using the LRP explanation technique. LRP explanations reveal that concepts do not map to distinct regions in the input space, similar to previous work with IG explanations. However, relevance from the final classification back to the concept vector shows the model has mapped these as expected for some CBM training methods. (Exceptionally, the sequential training and joint-without-sigmoid methods applies positive relevance to concepts not predicted as present and negative relevance to ones predicted as present.) We demonstrate the ability to calculate proportional concept contribution to final classifications. Both this and the saliency maps generated from the final classification to the concept vector should aid a human user to focus on the most important concepts and improve their mental model of the CBM error boundaries.

Future work will focus on instance-level concepts to analyse if the dataset alone can confine the model to learn distinct features in the input space or whether a new training method is required. A suitable dataset will require well-labelled concepts. With the challenge of accurate concept labelling, as we have seen with CUB, synthetic datasets may be a viable option. Selecting a dataset without a 1 to 1 correlation between the concepts and final classification, such as CelebA (Liu et al. 2015), has also yet to be explored for relevancy visualisations. In addition, a human study should be conducted to analyse the effectiveness and quality of the saliency maps and concept proportion contribution when used with CBMs.

Separate to saliency maps, it would be beneficial to remove concept(s) from an input to measure changes in concept and final classification predictions. This may be measured using techniques such as Remove and Retrain (Hooker et al. 2019) to avoid out of distribution samples affecting results.

Acknowledgements

We thank Liam Hiley for the support he provided in the early stages of this research. In particular, for guidance on adding LRP capabilities to the models.

This research is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and IBM UK via an Industrial CASE (ICASE) award.

References

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7): 1–46.

Craik, K. J. W. 1943. *The nature of explanation*. Oxford, England: University Press, Macmillan. ISBN 0674568826.

Grezl, F.; Karafiat, M.; Kontar, S.; and Cernocky, J. 2007. Probabilistic and Bottle-Neck Features for LVCSR of Meetings. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, IV–757–IV–760.

Halasz, F. G.; and Moran, T. P. 1983. Mental Models and Problem Solving in Using a Calculator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '83, 212–216. New York, NY, USA: Association for Computing Machinery. ISBN 0897911210.

Hooker, S.; Erhan, D.; Kindermans, P.-J.; and Kim, B. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 9737–9748. Curran Associates, Inc.

Johnson-Laird, P. N. 1986. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. USA: Harvard University Press. ISBN 0674568826.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept Bottleneck Models. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5338–5348. PMLR.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; and Weller, A. 2021. Do Concept Bottleneck Models Learn as Intended? arXiv:2105.04289.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. *Layer-Wise Relevance Propagation: An Overview*, 193–209. Springer International Publishing. ISBN 978-3-030-28953-9.

Norman, D. 1983. Some observations on mental models. *Human-Computer Interaction*, 241–244.

Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K.-R. 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109(3): 247–278.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. B.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3319–3328. JMLR.org.

Taylor, H.; Hiley, L.; Furby, J.; Preece, A.; and Braines, D. 2020. VADR: Discriminative Multimodal Explanations for Situational Understanding. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, 1–8.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Zhang, J.; Bargal, S. A.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2018. Top-Down Neural Attention by Excitation Backprop. *International Journal of Computer Vision*.